DOCUMENT RESUME

ED 467 376 TM 034 302

AUTHOR Glas, Cees A. W.; van der Linden, Wim J.

TITLE Computerized Adaptive Testing with Item Clones. Research Report.

INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

REPORT NO RR-01-10

PUB DATE 2001-00-00

NOTE 29p.

AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

PUB TYPE Reports - Research (143)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Adaptive Testing; Bayesian Statistics; *Computer Assisted Testing; *Costs; Item Response Theory; Maximum Likelihood Statistics; Test Construction; *Test Items

IDENTIFIERS *Cloning; Item Calibration; Item Parameters

ABSTRACT

To reduce the cost of item writing and to enhance the flexibility of item presentation, items can be generated by item-cloning techniques. An important consequence of cloning is that it may cause variability on the item parameters. Therefore, a multilevel item response model is presented in which it is assumed that the item parameters of a three-parameter logistic model describing response behavior are sampled from a multivariate normal distribution associated with a parent item. In this approach to item calibration, only distributions of item parameters are estimated. Therefore, the savings in item calibration costs for the item cloning model are potentially enormous. A marginal maximum likelihood and a Bayesian item calibration procedure are formulated. Further, a two-stage item selection procedure for computerized adaptive testing is presented. First, a set of items cloned from the same parent item is selected to be optimal at the ability estimate. Second, a random item from this set is administered. Simulation studies illustrate the accuracy of the item pool calibration and ability estimation procedures. An appendix describes Bayes model estimates for the item cloning model. (Contains 21 references.) (Author/SLD)

Computerized Adaptive Testing with Item Clones

Research Report 01-10

Cees A.W. Glas
W.J. van der Linden

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Computerized Adaptive Testing With Item Clones

Cees A.W. Glas

Wim J. van der Linden

Abstract

To reduce the cost of item writing and to enhance the flexibility of item presentation, items can be generated by item-cloning techniques. An important consequence of cloning is that it may cause variability on the item parameters. Therefore, a multilevel item response model is presented were it is assumed that the item parameters of a 3-parameter logistic model describing response behavior are sampled from a multivariate normal distribution associated with a parent item. In the present approach to item calibration, only distributions of item parameters are estimated. Therefore, the savings in item calibration costs for the item cloning model are potentially enormous. A marginal maximum likelihood and a Bayesian item calibration procedure are formulated. Further, a two-stage item selection procedure for computerized adaptive testing is presented: First, a set of items cloned from the same parent item is selected to be optimal at the ability estimate. Second, a random item from this set is administered. Simulation studies illustrate the accuracy of the item pool calibration and ability estimation procedures.

Keywords: computerized adaptive testing, item clones, item shells, multilevel item response theory, marginal maximum likelihood, Bayesian item selection.

Introduction

A major impediment to cost-effective implementation of computerized adaptive testing (CAT) is the amount of resources needed for item pool development. One of the solutions to the problem currently pursued is generating pools of items by using item-cloning techniques. Early pioneers of this idea were Bormuth (1970), Hively, Patterson and Page (1968) and Osburn (1968). Common to their approaches is a formal description of a set of "parent items" along with algorithms to derive a larger set of operational items from them. These parent items are known as "item forms", "item templates", or "item shells", whereas the items generated from them are now widely known as "item clones". We will use the term "parent item" to denote both the initial item and the set of clones generated from it.

Parent items may take the form of a syntactic description of a test item with one or more variable places for which substitution sets are specified. Clones are then generated by random draws from the substitution sets. In these "replacement set procedures" (Millman & Westman, 1989) the computer puts the answers to multiple-choice items in random order, picks distractors from a list of possible wrong answers, and, in numerical problems, substitutes random numbers in a specific spot in the item stem and adjusts the alternatives accordingly. Parent items may also consists of intact items from which clones are generated using transformation rules. Examples of such rules are linguistic rules that transform one verbal item into others, geometric rules that present objects from a different angle for spatial ability testing, transformations that allow one molecular structure to be derived from another in testing of knowledge of organic chemistry, or rules from proposition logic that generate items for testing of the ability in analytic reasoning. Comprehensive reviews of the technology of item cloning are given in Bejar (1993) and Roid and Haladyna (1982).

An important question is whether clones from the same parent item have comparable statistical characteristics. If they do, important savings in the costs of item pool calibration are possible, because it would then suffice to calibrate the characteristics of the parent only. In an extreme case, one might assume that the item parameters are constant over

the clones derived from the same parent. Empirical studies addressing this question are reported in, for example, Hively, Patterson and Page (1968), Macready (1983), Macready and Merwin (1973) and Meisner, Luecht and Reckase (1993). The general impression from these studies is that the variability between clones from the same parent is much smaller than between parents, but not small enough to justify the assumption of identical values. Of course, the size of the remaining variability depends on various factors, such as the type of knowledge or skill tested and the implementation of the item cloning technique.

The current paper is based on the expectation that attempts to improve item cloning techniques are desirable but that some degree of within-parent variability will always remain. The best way to deal with this variability is not to ignore it, but to model the distribution of the item parameters and allow for the uncertainty about their individual values when selecting the adaptive test.

A design for adaptive testing that fits in naturally with this approach is one with item selection based on stratified or two-staged sampling of items from the pool. In this sampling design, each item is selected in the following two steps: (1) A parent is selected from the pool with a set of clones that is optimal at the current ability estimate of the examinee; (2) A clone is randomly sampled from the set and administered to the examinee. This design capitalizes on the statistical advantage of administering tests with items adapted to the examinee's ability but, as will be discussed below, due the random sampling in the second step, also saves an important part of the resources needed for item calibration in regular CAT.

The proposed sampling design leads to a two-level item response theory (IRT) approach–with a lower level at which item clones are represented by a three-parameter logistic (3PL) model and a higher level at which the item parameters in this model are random with a (joint) distribution that represents within-parent variability. To capture between-parent variability in item parameter values, these distributions are allowed to vary in location and variance.

In the model below, the distributions of the item parameters for the parents are characterized by nine hyperparameters each. The values for these hyperparameters are

estimated from a data set where, for each examinee in the sample, one clone sampled from its parent. Because sampling is at random, the fact that the responses to the other clones from the same parent are missing can be ignored. Estimating nine hyperparameters per parent is the equivalent of calibrating three items under the 3PL model. Since item-cloning techniques easily lead to much large numbers of clones per parent, the savings in the resources needed when collecting calibration data are potentially enormous.

When selecting parent items in the first stage of the item-selection procedure, we have to cope with distributions rather than individual values for the item parameters. An obvious solution is to base selection of the parents on a Bayesian criterion with the distribution of the item parameters averaged out. The result is a reduction in the accuracy of ability estimation. Numerical examples of this reduction are shown in the empirical examples presented below, both relative to the case of regular CAT from a pool of individual items and a pool of cloned items calibrated under the regular 3PL model.

It is instructive to observe how the proposed type of adaptive testing can be viewed as an intermediate case of (1) classical domain-referenced testing under a binomial model (e.g., Lord & Novick, chap. 23) and (2) regular CAT from a pool of individual items. This type of CAT shares the idea of random item selection with the former and optimal selection at ability estimates with the latter. If all variability between the item-parameter values is within the parents, it is identical to domain-referenced testing. If all variability is between the parents, it is identical to CAT from a pool of individually written and calibrated items. However, if item cloning is effective, much smaller within-parent than between-parent variability is expected, and the proposed type of adaptive testing has efficiency close to regular CAT.

From a practical point of view it often is necessary to have test specialist review items generated by cloning algorithms before they are administered. The necessity of review becomes more crucial if (1) the domain of knowledge or skills contains socially sensitive material and (2) the algorithms can not be fully trusted. However, from a statistical point of view, it does not make much difference if in the second stage of item selection clones are drawn randomly from large sets of items generated and reviewed earlier that are stored

physically in computer memory or if they are generated on the fly by computer algorithms with a random seed. In either case the critical assumption of random sampling is met, and sampling is from approximately the same parameter distributions.

Model

Consider an item pool generated from parents $p = 1, ..., P$. The clones from parent p will be labeled $i_p = 1, ..., I_p$. The first-level model is the 3PL model, which describes the probability of success on item i_p as

$$p_{i_p}(\theta) \equiv \Pr\{X_{i_p} = 1\} \equiv c_{i_p} + (1 - c_{i_p})\frac{\exp[a_{i_p}(\theta - b_{i_p})]}{1 + \exp[a_{i_p}(\theta - b_{i_p})]}, \tag{1}$$

where X_{i_p} is the response variable for item i_p, with $X_{i_p} = 1$ for a correct and $X_{i_p} = 0$ for an incorrect response. The values of the item parameters $(a_{i_s}, b_{i_s}, c_{i_s})$ are realizations of a random vector. The second-level model describes the distribution of this vector through the transformation

$$\boldsymbol{\xi}_{i_p} = (\log a_{i_p}, b_{i_p}, \text{logit } c_{i_p}) \tag{2}$$

with a multivariate normal distribution

$$\boldsymbol{\xi}_{i_p} \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \tag{3}$$

where $\boldsymbol{\mu}_p$ is the vector with the mean values of the item parameters for parent p and $\boldsymbol{\Sigma}_p$ their covariance matrix. The transformation in (2) is introduced to give the item parameters scales for which the assumption of multivariate normality in (3) is reasonable.

In the calibration and item selection procedures below, we will assume that θ has a standard normal prior distribution, that is,

$$\theta \sim N(0, 1). \tag{4}$$

This assumption holds if j is from a population of exchangeable examinees with a normal distribution of abilities.

The model presented in (1)-(4) has several relatives. The multilevel IRT models for testlets in Bradlow, Wainer & Wang (1999) and Wainer, Bradlow, and Zu (2000) differ fro the present model in having a random component for difficulty parameter b_i but fixed parameters a_i and c_i. The random component is used to allow for dependence between responses to fixed items in the same testlet. Because our items are randomly sampled from parents, all item parameters need to be random and dependence between responses to items from the same parent is captured by the covariance matrix in (3). The present model also differ from the one in Albers, Does, Imbos and Jansen (1989) and Janssen, Tuerlinckx, Meulders and de Boeck (2000) who also assume item sampling but model the process by a version of the 1PL model with a random difficulty parameter.

Item Pool Calibration

In the present approach, item pool calibration amounts to estimation of the values for each parent of the hyperparameters in the distribution in (3). It is assumed that these parameters are stacked in a vector $\eta \equiv (\mu_1, \Sigma_1, ..., \mu_P, \Sigma_P)$. The values of these parameters can be estimated by the methods of marginal maximum likelihood (MML) or Bayes modal estimation (MAP).

The response vector of examinee j is denoted as $\mathbf{x}_j \equiv (x_{i_p j}) \equiv (x_{i_1 j,...,} x_{i_P j})$, where i_p is item clone i randomly drawn from parent p. As already noted, estimation of vector η is from a data set with for each examinee j the responses to one item clone sampled from its parent. Because the responses to the other item clones are missing at random, they can be ignored. In practice, the adaptive nature of the test will also involve sets of calibration data with examinees missing parents. These data are missing at random too. However, to save unnecessary complexity, our notation will not make this incompleteness explicit.

MML Calibration

In MML estimation, a distinction in made between structural and nuisance parameters. The structural parameters are estimated from a log-likelihood marginalized with respect to the nuisance parameters. In the present case, the structural parameters are in the vector η, whereas the nuisance parameters are the ability parameters θ and the random item parameters ξ_{i_p}. These nuisance parameters are supposed to be stacked in vectors θ and ξ, respectively.

The marginal probability of observing response pattern $\mathbf{x}_j$ is given by

$$p(\mathbf{x}_j;\ \eta) = \int \dots \int p(\mathbf{x}_j \mid \theta, \xi) f(\xi, \theta|\eta) d\xi d\theta \tag{5}$$

$$= \int \dots \int \prod_p p(\mathbf{x}_{i_p j} \mid \theta, \xi_{i_p}) h(\xi_{i_p}|\mu_p, \Sigma_p) \phi(\theta) d\xi_{i_p} d\theta \tag{6}$$

$$= \int \left[\prod_p \int \dots \int p(\mathbf{x}_{i_p j} \mid \theta, \xi_{i_p}) h(\xi_{i_p}|\mu_p, \Sigma_p) d\xi_{i_p} \right] \phi(\theta) d\theta. \tag{7}$$

The marginal log-likelihood of η is given by

$$\log L(\eta;\mathbf{x}) = \sum_j \log p(\mathbf{x}_j;\eta). \tag{8}$$

The marginal likelihood equations for η can be easily derived using Fisher's identity (Efron, 1977; Louis 1982). The first-order derivatives with respect to η can be written as

$$\frac{\partial}{\partial \eta} \log L(\eta;\mathbf{x}) = \sum_j E\left(\frac{\partial}{\partial \eta} \log f_j(\xi, \theta_j|\eta) \mid \mathbf{x}_j, \eta\right) = \mathbf{0}, \tag{9}$$

where $\log f_j(\xi, \theta_j|\eta)$ is the so-called "complete data" log-likelihood

$$\log f(\xi_{i_p}, \theta|\eta) =$$

$$\sum_p \log p(\mathbf{x}_{i_p j} \mid \theta, \xi_{i_p}) + \sum_p \log p(\xi_{i_p}|\eta) + \log \phi(\theta),$$

and the expectation is with respect to the conditional posterior density for the nuisance parameters, that is, with respect to

$$p(\boldsymbol{\xi}_{i_p}, \theta \mid \mathbf{x}_j, \boldsymbol{\eta}) \propto \prod_p p(\mathbf{x}_{i_p j} \mid \theta, \boldsymbol{\xi}_{i_p}) p(\boldsymbol{\xi}_{i_p} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \phi(\theta). \tag{10}$$

It follows that the likelihood equations are given by

$$\mu_{pu} = \sum_j E(\xi_{pu} \mid \mathbf{x}_j, \boldsymbol{\eta}), \tag{11}$$

$$\sigma^2_{pu} = \sum_j E(\xi^2_{pu} \mid \mathbf{x}_j, \boldsymbol{\eta}) - \mu^2_{pu}, \tag{12}$$

and

$$\sigma_{puv} = \sum_j E(\xi_{pu}\xi_{pv}\mathbf{x}_j, \boldsymbol{\eta}) - \mu_{pu}\mu_{pv}, \tag{13}$$

where indices u and $v \neq u$ denote the uth and vth element in the parameter vectors. These equations can be solved using an EM or Newton-Raphson algorithm.

Computation of the standard errors of the parameters estimates is a straightforward generalization of the method for the 3PL model presented in Glas (2000). These estimates are found upon inverting the approximate information matrix

$$\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta}) \approx \sum_j E\left[\frac{\partial}{\partial \boldsymbol{\eta}} \log f_j(\boldsymbol{\xi}, \theta_j | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \mid \mathbf{x}_j, \boldsymbol{\eta}\right] E\left[\frac{\partial}{\partial \boldsymbol{\eta}} \log f_j(\boldsymbol{\xi}, \theta_j | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \mid \mathbf{x}_j, \boldsymbol{\eta}\right]'.$$

Bayes Modal Calibration

The use of Bayes modal estimation can be motivated by the fact that the parameters in the 3PL model are sometimes hard to estimate because they are poorly determined by the available data. In such instances, the behavior of the item response functions over the region of the ability scale where the respondents are located can be described by different

combinations of parameter values. As a result, the estimates of the parameters in the 3PL model are highly correlated. Adding a covariance matrix for every parent may worsen the identifiability of the model for such data sets.

To obtain "reasonable", finite estimates, Mislevy (1986) considered a number of Bayesian approaches. Each of them entails the introduction of prior distributions on the item parameters. Parameter estimates are computed maximizing the log-posterior density of η, which is proportional to $\log L(\eta;\ \mathbf{x}) + \log p(\eta \mid \zeta) + \log p(\zeta)$, where $p(\eta \mid \zeta)$ is the prior density of the η, characterized by parameters ζ, which in turn follow a density $p(\zeta)$. In one approach, the prior distribution $p(\eta \mid \zeta)$ is postulated by fixed the item calibrator; in another, often labeled empirical Bayes, the parameters of the prior distribution are estimated along with the other parameters, for example, as the modes of their posterior distribution. In our case, the second approach is formally identical to Bayes modal or maximum a posterior (MAP) estimation of the parent parameters, albeit that the estimates have to be found for all parents simultaneously. The approach involves a change of the likelihood equations to $\partial \log L(\eta \mid \mathbf{x})\partial\eta + \partial \log p(\eta \mid \mathbf{x})\partial\eta = \mathbf{0}$, while simultaneously the equations $\partial \log p(\eta \mid \zeta)/\partial\zeta + \partial \log p(\zeta)/\partial\zeta = \mathbf{0}$ must be solved. An outline of the procedure for the current item cloning model is given in Appendix A.

Discussion

The assumption that all respondents are drawn from one population can be replaced by the assumption of multiple populations of respondents each with a normal ability distribution indexed by a unique mean and variance parameter. Bock and Zimowski (1997) point out that this generalization, together with the possibility of analyzing incomplete item-calibration designs, provides a unified approach to such problems as differential item functioning, item parameter drift, non-equivalent groups equating, vertical equating and matrix-sampled educational assessment. Though not illustrated here, calibration under the item-cloning model can also be extended to fit this framework.

Adaptive Selection of Parent Items

Our initial estimate of the ability of examinee j is the prior distribution in (4), which has a density denoted as $\phi(\theta)$. Suppose parents $1, ..., k-1$ have been selected. For each parent a clone has been administered, the responses to which are denoted by a vector $\mathbf{x}_j^{(k-1)} \equiv (x_{j1}, ..., x_{j(k-1)})$. Then the posterior distribution of θ given $\mathbf{x}_j^{(k-1)}$ is

$$p(\theta \mid \mathbf{x}_j^{(k-1)}) \propto \phi(\theta) \prod_{p=1}^{k-1} \int p(x_{jp} \mid \theta, \boldsymbol{\xi}_p) p(\boldsymbol{\xi}_p | \boldsymbol{\mu}_p, \Sigma_p) d\boldsymbol{\xi}_p. \tag{14}$$

The variance of this posterior distribution is denoted as $\text{Var}(\theta|\mathbf{x}_j^{(k-1)})$.

The kth parent should be selected to be optimal at this posterior distribution. Several Bayesian criteria of optimality have been suggested; for studies of several old and new criteria, see van der Linden (1998) and van der Linden and Pashley (2000). The one used in the computer simulations below is the criterion of minimum expected posterior variance adapted for use with the item-cloning model. It selects the kth parent to have minimum posterior variance averaged both over the set of clones associated with the parent and the responses to the clones predicted from the examinees current ability estimate.

If parent p in the pool would be selected as the kth parent in the test, the posterior predictive distribution of the response of examinee j to a random item from this parent given the previous responses $\mathbf{x}_j^{(k-1)}$ is given by

$$f(x_{jp_k} \mid \mathbf{x}_j^{(k-1)}) = \int \left[\int p(x_{jp_k} \mid \theta, \boldsymbol{\xi}_{p_k}) p(\boldsymbol{\xi}_{p_k} | \boldsymbol{\mu}_{p_k}, \Sigma_{p_k}) d\boldsymbol{\xi}_{p_k} \right] p(\theta \mid \mathbf{x}_j^{(k-1)}) d\theta. \tag{15}$$

Note that the probability of the response is first averaged over the distribution of the item parameters for parent p_k and then over the posterior distribution of the ability of the examinee.

The two possible responses lead to updates of the posterior variance which we denote as $\text{Var}(\theta|\mathbf{x}_j^{(k-1)}, X_{jp_k} = 0)$ and $\text{Var}(\theta|\mathbf{x}_j^{(k-1)}, X_{jp_k} = 1)$. The proposed criterion for the

selection of the kth parent is the expected value of this update. That is,

$$p_k \equiv \arg\min_r \Big\{ \mathrm{Var}(\theta|\mathbf{x}_j^{(k-1)}, X_{jr_k} = 0) f(0 \mid \mathbf{x}_j^{(k-1)}) + \mathrm{Var}(\theta|\mathbf{x}_j^{(k-1)}, X_{jr_k} = 1) f(1 \mid \mathbf{x}_j^{(k-1)}); r \in R_k \Big\}, \tag{16}$$

where R_k is the set of parents in the pool from which the kth parent is chosen.

Simulation Studies

Two simulation studies were conducted. One study was to address the accuracy of the MML calibration procedure for the item cloning model in (11)-(13) under a variety of conditions. The other to address the accuracy of the ability estimator from the item selection procedure based on the criterion in (16) under the same conditions.

Three different types of CAT were studied, namely CAT from a pool of:

(1) cloned items calibrated and administered under the item cloning model;

(2) individual items calibrated and administered under the regular 3PL model;

(3) cloned items calibrated and administered under the regular 3PL model.

The comparison between Type1 and Type 2 helps us to identify the potential loss in accuracy due to second-stage item sampling and the presence of random item parameters in the item cloning model. The comparison between Type 1 and Type 3 shows us the statistical consequences of adaptive testing from a pool of cloned items under a conventional model that ignores the dependences between responses to items cloned from the same parent. These dependences are created by the fact that such items share certain structural features and attributes. The regular 3PL model in Type 3 CAT does not allow for such dependences, whereas the multilevel IRT model in (1)-(3) for Type 1 CAT does.

Items

Because the composition of the item pool can have a substantial impact on item calibration and ability estimation results in CAT, the items used in each of the three types of CAT were generated using a common multivariate normal distribution for the

(transformed) item parameters $(\log a, b, \text{logit}\, c)$, with mean

$$\mu_0 = (.0, .0, \text{logit}(.2)) \tag{17}$$

and covariance matrix

$$\Sigma_0 = \begin{bmatrix} 0.20 & 0.05 & -0.05 \\ 0.05 & 1.00 & 0.10 \\ -0.05 & 0.10 & 0.10 \end{bmatrix}. \tag{18}$$

Item pools with a cloning structure were obtained by sampling the values for the vector of means of the distribution of the item parameters for each parent in (3), μ_p, from (17)-(18). The covariance matrices of these distributions were all equated to the matrix in (18); that is, Σ_p was set equal to Σ_0 for all p. Pools with individual items were obtained sampling their true item parameter values from the distribution in (17)-(18). To approximate the composition of the previous type of pool as closely as possible, the pools were refreshed for each replication.

Calibration

In this simulation study, the following additional variables were manipulated:

(1) test length: n=20, 30 and 40 items;

(2) sample size: N=100, 400 and 1,000 examinees.

For each condition, N examinees were simulated, drawing random values for θ from the standard normal distribution. The mean absolute error in the estimates of the parameter in the item cloning model (Type 1 CAT) or the 3PL model used to calibrate the item pools (Type 2 CAT) are shown in Table 1.

Insert Table 1 about here

The pattern in the errors for the two models are approximately the same. As expected, the errors decreased both in the size of the sample and the length of the test, and generally larger errors were obtained for the discrimination than for the difficulty parameters. The

last three columns show the differences in mean absolute error between the parameters estimates for the two models. The differences between the errors in the estimates of the guessing parameter are negligible. The differences between the errors in the estimates of the difficulty and discrimination parameters are small but, as expected, systematically in favor of those for the regular CAT model. (Observe that these two sets of parameters are on identical scales but have distributions true values that show random differences. For the given item pool size, the effect on the comparison in Table 1 can be assumed to be negligible, though.)

In Table 2, the same comparison is made for the parameters estimates for a pool of cloned items calibrated under the item cloning model in this paper (Type 1 CAT) and a regular 3PL model that ignores the item cloning structure (Type 3 CAT). The differences are generally larger than in the previous comparison.

Insert Table 2 about here

The covariance matrix in (18) could be estimated only for calibration under the item cloning model. The mean absolute estimation errors are given in Table 3. Observe that the errors in the estimates of the variances decrease both in the sample size and the test length but that the decrease is negligible for the estimates of the covariances. Generally, but not unanticipated, estimation of the covariance matrix appeared to be much less accurate than estimation of the vector of means of the parameters in the item cloning model.

Insert Table 3 about here

Ability Estimation

The same three types of CAT as in the calibration study were studied. The size of the pool was always equal to 400. The final ability estimates in Type 1 CAT were calculated as the expected value of the posterior distribution (EAP estimate) in (14). In the other

two types of CAT, EAP estimates under the regular 3PL model with the prior in (4) were calculated.

The following additional variables were manipulated:

(1) test length: n=20, 30 and 40 items.

(2) true ability value: θ= -2.0. -1.0, 0.0, 1.0, and 2.0, and $\theta \sim N(0, 1)$.

For each condition, 400 examinees were simulated. The item parameters were redrawn for every simulee. The mean absolute errors in the ability estimates are shown in Table 4. The comparison between the errors for Type 1 and Type 2 CAT shows the price in efficiency to be paid for item cloning with second-stage random sampling of clones from parent items. The differences were negligible for θ values close to zero but increased toward the tails of the θ distribution. This change is due to the use of the standard normal prior in (4) which favors item selection near $\theta = 0$ at the beginning of the test for both types of CAT. The comparison between Type 3 and Type 1 CAT shows the additional loss of accuracy if the dependencies between responses to items cloned from the same parent is not modeled. These differences were negligible for θ values close to zero but again increased toward the tails of the θ distribution. The average error across sampling of examinees from a standard normal population showed the same pattern but with smaller values. Also, both series of differences showed a tendency to decrease in the length of the test, albeit the tendency was smaller for the types of CAT with item cloning than for regular CAT.

Insert Table 4 about here

Conclusion

The advantage of CAT with item cloning is a potentially large reduction in the resources needed for item pool development. The price to be paid for this advantage is a reduction in the accuracy of the ability estimates. For the typical test length in the current adaptive testing programs of $n = 30$, the decrease in the average accuracy of

ability estimation across a normal population of examinees was slightly over 10% for the multilevel model in this paper. The decrease can easily be compensated by added 2-4 items to the test. It is left to the testing agent to decide if the trade-off by the reduction in item pool development costs and test length is advantageous.

References

Albers, W., Does, R. J. M. M., Imbos, Tj., & Janssen, M. P. E. (1989). A stochastic growth model applied to repeated test of academic knowledge. *Psychometrika, 54*, 451-466.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer-Verlag.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago, IL: University of Chicago Press.

Bradlow, E. T., Wainer, H., Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. P. Demster, N. M. Laird and D. B. Rubin). *Journal of the Royal Statistical Society* (Series B), *39*, 1-38.

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 183-199). Norwell, MA: Kluwer Academic Publishers.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A 'universe-defined" system of arithmetic achievement items. *Journal of Educational Measurement, 5*, 275-290.

Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal od Educational and Behavioral Statistics, 25*, 285-306.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226-233.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addision-Wesley.

Macready, G. B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. *Applied Psychological Measurement, 7,* 149-157.

Macready, G. B., & Merwin, J. C. (1973). Homogeneity witin item forms in domain referenced testing. *Educational and Psychological Measurement, 33, 351-360.*

Meisner, R., Luecht, R. M., Reckase, M. D. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms* (ACT Reserach Report Series Nop. 93-9). Iowa City, IA: ACT, Inc.

Millman, J., & Westman, R.S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement, 26,* 177-190.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Pychometrika, 51,* 177-195.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurements, 28,* 95-104.

Roid, G., & Haladyna, T. 91982). *A technology for test-item writing*. New York: Academic Press.

van der Linden, W.J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika, 63,* 201-216.

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Norwell, MA: Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-bsed adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Norwell, MA: Kluwer Academic Publishers.

Appendix A: Bayes Model Estimates for the Item Cloning Model

The marginal probability of observing response pattern $\mathbf{x}_j$ is enhanced with a conjugate prior $p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. The conjugate prior distribution for $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ is a product of a normal and an inverse-Wishart distribution (see, for instance, Box & Tiao, 1973). The marginal probability of examinee j's response vector now becomes

$$p(\mathbf{x}_j;\ \boldsymbol{\eta}) = \int \cdots \int \prod_p p(\mathbf{x}_j \mid \theta, \boldsymbol{\xi}_{jp}) p(\boldsymbol{\xi}_{jp} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \phi(\theta) d\boldsymbol{\xi}_{jp} d\theta. \tag{A.1}$$

Consider the complete data specification

$$p(\mathbf{x}, \boldsymbol{\xi}, \theta \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_j \prod_p p(\mathbf{x}_{jp} \mid \theta, \boldsymbol{\xi}_p) p(\boldsymbol{\xi}_{jp} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \phi(\theta). \tag{19}$$

The factors

$$\prod_j \prod_p p(\boldsymbol{\xi}_{jp} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

entail a normal model with a normal-inverse-Wishart prior, with parameters, $\boldsymbol{\mu}_0$and $\boldsymbol{\Sigma}_0$, υ_0 the degrees of freedom for the prior of $\boldsymbol{\Sigma}_p$ and κ_0 the degrees of freedom for $\boldsymbol{\mu}_0$. Then the posterior is also inverse-Wishart distributed with parameters

$$\begin{aligned}
\boldsymbol{\mu}_s &= \tfrac{n}{\kappa_0+n}\bar{\boldsymbol{\xi}}_s + \tfrac{\kappa_0}{\kappa_0+n}\boldsymbol{\mu}_0 \\
\upsilon &= \upsilon_0 + n \\
\kappa &= \kappa_0 + n \\
\boldsymbol{\Sigma}_p &= \mathbf{S}_p + \tfrac{\kappa_0 n}{\kappa_0+n}(\bar{\boldsymbol{\xi}}_p - \boldsymbol{\mu}_0)(\bar{\boldsymbol{\xi}}_p - \boldsymbol{\mu}_0)^T + \boldsymbol{\Sigma}_o,
\end{aligned}$$

where $\mathbf{S}_p = \sum_{j=1}^{n} (\boldsymbol{\xi}_p - \bar{\boldsymbol{\xi}}_p)(\boldsymbol{\xi}_p - \bar{\boldsymbol{\xi}}_p)^T$.

As can be verified in (9), the likelihood equations are the posterior expectations of the first-order derivatives of the complete data likelihood. Analogous to (11)-(13), we

now have

$$\mu_p = \frac{1}{\kappa_0 + n} \sum_j E(\xi_s \mid \mathbf{x}_j, \eta) + \frac{\kappa_0}{\kappa_0 + n} \mu_0, \tag{A.3}$$

and

$$\Sigma_p = \sum_j E\left[(\xi_p - \bar{\xi}_p)(\xi_p - \bar{\xi}_p)^T \mid \mathbf{x}_j, \eta\right] + \frac{\kappa_0}{\kappa_0 + n}(\bar{\xi}_p - \mu_0)(\bar{\xi}_p - \mu_0)^T + \Sigma_o, \tag{A.4}$$

with

$$\bar{\xi}_p = \sum_j E(\xi \mid \mathbf{x}_j, \eta).$$

Table 1
Mean absolute error in item parameter estimates

		Type 1			Type 2			1-2		
n	N	α	β	γ	α	β	γ	α	β	γ
20	100	.312	.395	.061	.304	.385	.060	.008	.010	.001
	400	.238	.292	.056	.219	.278	.057	.019	.014	-.001
	1000	.201	.241	.051	.172	.222	.051	.029	.019	.000
30	100	.306	.384	.059	.295	.369	.059	.011	.015	.000
	400	.228	.281	.054	.213	.261	.054	.015	.020	.000
	1000	.189	.251	.052	.165	.230	.052	.024	.021	.000
40	100	.299	.384	.060	.287	.374	.060	.012	.010	.000
	400	.222	.286	.055	.202	.264	.055	.020	.022	.000
	1000	.189	.229	.051	.161	.219	.050	.028	.010	.001

Table 2
Mean absolute error in item parameter estimates

		Type 3			Type 1			3-1		
n	N	α	β	γ	α	β	γ	α	β	γ
20	100	.353	.409	.061	.312	.395	.061	.041	.014	.000
	400	.277	.310	.054	.238	.292	.056	.039	.018	-.002
	1000	.244	.275	.051	.201	.241	.051	.043	.034	.000
30	100	.321	.407	.058	.306	.384	.059	.015	.023	-.001
	400	.259	.303	.054	.228	.281	.054	.029	.022	.000
	1000	.241	.277	.052	.189	.251	.052	.052	.026	.000
40	100	.321	.400	.057	.299	.384	.060	.022	.016	-.003
	400	.257	.303	.054	.222	.286	.055	0.35	.017	-.001
	1000	.238	.277	.052	.189	.229	.051	.049	.048	.001

Table 3
Mean absolute error for estimates of item covariance matrix

S	N	$\sigma_{\log\alpha}$	σ_{β}	$\sigma_{\log it\gamma}$	$\sigma_{\log\alpha\beta}$	$\sigma_{\log\alpha\, \log it\gamma}$	$\sigma_{\beta logit\gamma}$
20	100	.040	.278	.0070	.149	.024	.122
	400	.028	.223	.0066	.136	.017	.125
	1000	.026	.218	.0053	.132	.017	.122
30	100	.042	.241	.0070	.141	.024	.123
	400	.027	.223	.0068	.132	.017	.122
	1000	.027	.215	.0050	.137	.017	.126
40	100	.034	.275	.0058	.134	.020	.110
	400	.026	.206	.0055	.116	.016	.105
	1000	.025	.207	.0050	.116	.016	.107

Table 4
Mean absolute error in ability estimates

n	Type of CAT	-2.0	-1.0	θ 0.0	1.0	2.0	Standard Normal
20	1	.560	.285	0.263	.268	.421	.291
	2	.438	.256	0.257	.202	.348	.282
	3	.557	.292	0.264	.248	.468	.290
30	1	.476	.285	.261	.225	.365	.260
	2	.364	.240	.257	.153	.275	.230
	3	.489	.279	.256	.207	.403	.258
40	1	.436	.265	0.255	.175	.307	.223
	2	.332	.219	0.255	.132	.248	.204
	3	.453	.247	0.264	.152	.341	.234

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 L.S. Sotaridona & R.R. Meijer, *Two New Statistics to Detect Answer Copying*

RR-01-06 L.S. Sotaridona & R.R. Meijer, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 I. Hendrawan, C.A.W. Glas, & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

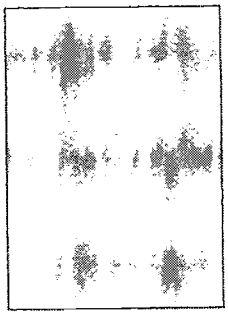
RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").